Electronic Supplement to Xia, Wang et al.

Table of contents

R is a free statistical software package that can be downloaded from www.r-project.org. Documentation can also be found on this site. This appendix contains the necessary R code for performing the G test, the two-sample t test, the *q*-value calculation, the q-q plot comparison of simulated G statistics with Chi-square distribution with 1 degree of freedom, the LOWESS curves for upper and lower boundary of the replicate runs of PG_nm, etc. Comments about the commands are included, they are the written after the ## signs.

Part 1. Analysis of spectral counts, G test and *q*-value calculation:

```
counts = read.table(file.choose(),header=T)
## Read in the spectral counts data. This will open a browser window to locate the file
## containing the data. Data should be in a tab-delineated text file in the form of:
##
##      NL.ScPP      Sc.nm  ScRatio1
##      484          858          0.56
##      167          166          1.01

f.norm = counts$Sc.nm    ## frequency in PG_nm
f.PP = counts$NL.ScPP    ## frequency in PG_PP

G = 2 * ( f.norm*log(f.norm/( (f.norm + f.PP)/2 ) + f.PP*log(f.PP/( (f.norm + f.PP)/2 ) )
## calculates the G-statistic

pvals = 1 - pchisq(G,df=1)   ## calculates p-values

install.packages("qvalue")
library(qvalue)
## install the q-value software (computer must be connected to the internet)

qobj = qvalue(pvals)
qobj$qvalue
## calculate the q-values
```

Part 2. Analysis of peptide signal intensities, two-sample t test and q value calculation:

```
intensities = read.table(file.choose(),header=T)
## Read in the peptide signal intensities data. This will open a browser window to locate
## the file containing the data.  Data should be in a tab-delineated with one line for each
## protein containing the sum in the PP state, the sum in the normal state, the number
## summed in the PP state, and the number summed in the normal state, and the standard
## deviations in each of the states:
##
##      PP      Normal      n.PP    m.Normal    sd.PP   sd.Normal
##      8763    8726        10      15          34.4            273.6
##      9328    2734        8       16          23.1            324.2

X = intensities$Normal
Y = intensities$PP

n = intensities$n.Normal
m = intensities$m.PP

s.X = intensities$sd.Normal
s.Y = intensities$sd.PP

z = (X-Y)/sqrt(n*(s.X^2) + m*(s.Y^2))
## calculates the z-statistic

pvals = 2*(1 – pnorm(abs(z),0,1))   ## calculates p-values

install.packages(“qvalue”)
library(qvalue)
## install the q-value software (computer must be connected to the internet)

qobj = qvalue(pvals)
qobj$qvalue
## calculate the q-values
```

Part 3. Simulated G statistics from two binomial distributions and the q-q plot for comparing it with Chi-square distribution with 1 degree of freedom:

```
p=0.5
## set the probability for binomial distribution

n=1000
## set the sample size

G = c()
## use G as a vector to contain the G values

for (i in 1:n)    {

        f1 = rbinom(1,n, p)
        f2 = rbinom(1,n, p)
        G[i] = 2 * (f1*log(f1/((f1+f2)/2))+f2 * log(f2/((f1+f2)/2)))
}
## Use two binomial distribution to generate simulated G statistics

qqplot(G,rchisq(n,1),xlab = "quantile of G", ylab = "quantile of chi-square")
## draw qq plot, x is quantile of G, y is quantile of chi-square

lines(c(0:floor(max(G))), c(0:floor(max(G))))
## draw a line of y = x, the length of the line depends on floor(max(G))
```

Part 4. Generate scatter plot of protein level spectral count ratios of replicate runs of PG_nm and the summed spectral counts, add LOWESS curves to the upper and lower boundary. Generate scatter plot for protein level spectral count ratios of PG_PP/PG_nm and the summed spectral counts, superimpose the LOWESS curves from the scatter plot of PG_nm replicate runs onto the plot and calculate the numbers of proteins those are outside the curves and generate a list which contains those ORF numbers:

```
## First, plot PG_nm replicate runs data and draw LOWESS curves

test = read.table(file.choose(), header =T)
## Read in the data. This will open a browser window to locate the file containing the
## data.  Data should be in a tab-delineated with one line for each protein containing the
## normalized spectral count ratio, the summed spectral counts:
##
##      Protein         Log2(TotalCounts)    Log2(NL_Ratio)
##      PG1545          9.8                  -0.06
##      PG0521          7.43                 0.06

ratio = test$Log2.NL_Ratio.
count = test$Log2.TotalCounts.

breaks = hist(count, plot=F, nclass = 25)$breaks

nbreaks = length(breaks)

max.index = c()
min.index = c()

for (i in 1:(nbreaks -1)){

        max.index[i] = which.max(ratio*(count >= breaks [i] & count < breaks[i+1]))
        min.index[i] = which.min(ratio*(count >= breaks[i] & count < breaks[i+1]))
}

plot(count,ratio, ylim = c(-8,8), xlim = c(0,14))


lines(lowess(count[max.index], ratio[max.index]))
lines(lowess(count[min.index], ratio[min.index]))
```

```
## Second, plot PG_PP/PG_nm data and draw LOWESS curves

test2 = read.table(file.choose(), header =T)

ratio2 = test2$Log2.NL_Ratios.
count2 = test2$Log2.TotalCounts.

breaks2 = hist(count2, plot=F, nclass = 25)$breaks

nbreaks2 = length(breaks2)

plot(count2,ratio2, ylim = c(-8,8),xlab="Log2 of sum of spectral counts from PG_PP and
PG_nm", ylab = "Log2 of PG_PP/PG_nm spectral counts")

lines(lowess(count[max.index], ratio[max.index]))
lines(lowess(count[min.index], ratio[min.index]))



## Third, calculate how many above LOWESS curve and how many below

above.LOWESS = c()
below.LOWESS = c()

for (i in 1:(nbreaks2 -1)){

        above.LOWESS[i] =
sum(ratio2*(count2>=breaks2[i]&count2<breaks2[i+1])>=lowess(count[max.index],
ratio[max.index])$y[i+1])
        below.LOWESS[i] =
sum(ratio2*(count2>=breaks2[i]&count2<breaks2[i+1])<=lowess(count[min.index],
ratio[min.index])$y[i+1])
}

sum(above.LOWESS)
sum(below.LOWESS)



## Fourth, to pull out the proteins those are outside the LOWESS curves


proteins = test2$Protein
```

```
above.LOWESS.proteins = c()
below.LOWESS.proteins = c()


for (i in 1:(nbreaks2 -1)){

        above.LOWESS.proteins =
proteins[ratio2*(count2>=breaks2[i]&count2<breaks2[i+1])>=lowess(count[max.index],
ratio[max.index])$y[i+1]]
        below.LOWESS.proteins =
proteins[ratio2*(count2>=breaks2[i]&count2<breaks2[i+1])<=lowess(count[min.index],
ratio[min.index])$y[i+1]]

        write.table(above.LOWESS.proteins, file = "C:\\Temp\\aboveproteinlist.txt",
append = T)

        write.table(below.LOWESS.proteins, file = "C:\\Temp\\belowproteilist.txt",
append = T)
}
```

Electronic Supplement Table S1

Ratios for PG_PP/PG_nm and their statistics for each of the 20 proteins shown to be significant using both control populations of PG. Under-expressed proteins are shaded in gray.

Protein	Log2 ScRatio1	*q*-Val	Log2 ScRatio3	SD	n	Log2 PiRatio1	*q*-Val	Log2 PiRatio3	SD	n	Expression Direction
P13793|FMA_PORGI	-3.64	0	-2.45	1.34	11	-3.94	0	-3.34	1.92	26	↓
PG0121	2.21	0	1.80	1.55	3	1.53	2.26E-07	4.77	3.41	5	↑
PG0293	2.47	0	1.87	1.34	9	2.49	0	2.78	2.04	15	↑
PG0377	2.24	0	2.44	1.39	9	2.63	0	4.52	3.42	21	↑
PG0378	3.55	0	3.80	1.22	9	2.49	1.46E-10	5.74	2.96	17	↑
PG0386	2.73	0	1.75	1.34	2	3.67	2.00E-07	6.00	3.82	5	↑
PG0449	2.66	0	1.75	0.85	5	3.14	0	5.72	2.80	25	↑
PG0776	1.66	0	1.31	0.36	5	1.66	3.95E-08	1.80	1.53	14	↑
PG1006	-3.64	0	-2.45	1.48	5	-5.29	0	-3.97	1.41	31	↓
PG1082	1.96	0	1.49	0.87	8	1.10	9.03E-05	2.37	2.29	16	↑
PG1085	3.32	0	3.49	0.84	3	2.57	7.05E-05	6.63	3.98	5	↑
PG1279	3.11	0	3.02	1.75	9	2.97	0	6.07	2.94	17	↑
PG1341	1.98	0	1.62	0.53	4	2.20	7.80E-11	3.33	2.34	9	↑
PG1551	-0.97	0	-1.10	1.00	9	-1.07	2.17E-04	-1.13	0.54	9	↓
PG1729	1.28	0	1.20	0.48	5	1.08	7.61E-06	1.98	1.54	11	↑
PG1809	3.03	0	3.13	1.01	7	2.26	3.44E-11	4.97	1.59	8	↑
PG1928	1.08	7.97E-12	2.11	1.57	3	1.38	3.29E-07	2.54	0.97	4	↑
PG1940	3.35	0	3.34	1.32	22	2.62	0	6.15	2.91	38	↑
PG2130	-5.64	0	-2.62	0.62	7	-6.89	1.80E-06	-6.15	1.92	9	↓
PG2168	-4.06	0	-2.26	0.91	6	-6.55	1.83E-07	-5.18	2.82	11	↓

All four relative quantitation methods gave the same direction of change for the 20 proteins common to both comparisons (see Table 1 in the printed text) that fit the strict criteria: *q*-values less than 0.05, $\log_2$ of peptide level ratios + SD < 0 or $\log_2$ of peptide level ratios - SD > 0. Many proteins in Table S1 and Table S2 have *q*- values shown as zero. The smallest positive floating-point number with double precision in R is 2.220446 Exp-16. Any number less than this value will be shown as zero.

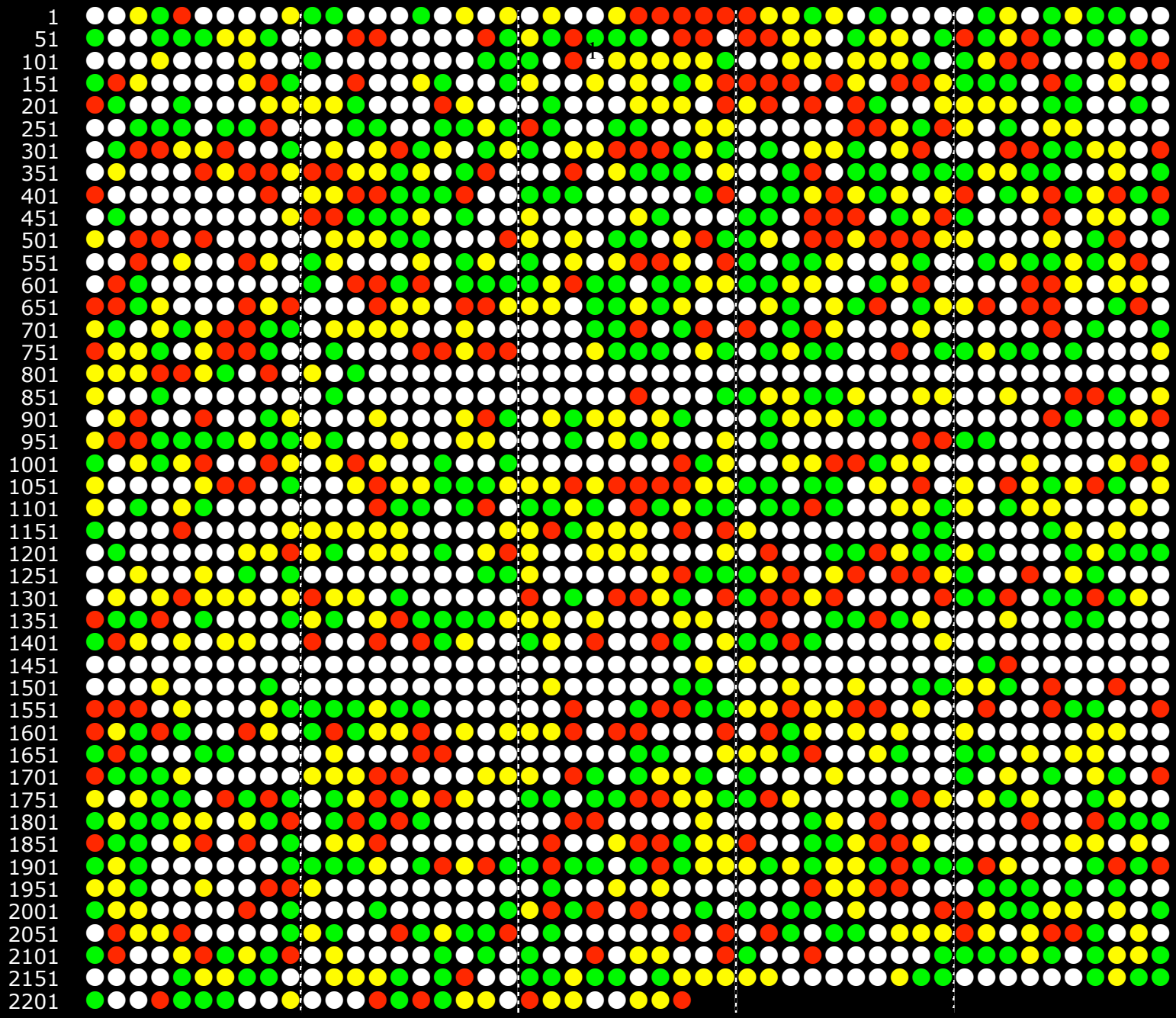
Electronic Supplement Table S2

Ratios for PG_PP/PG_PPC and their statistics for each of the 20 proteins shown to be significant using both control populations of PG. Under-expressed proteins are shaded in gray.

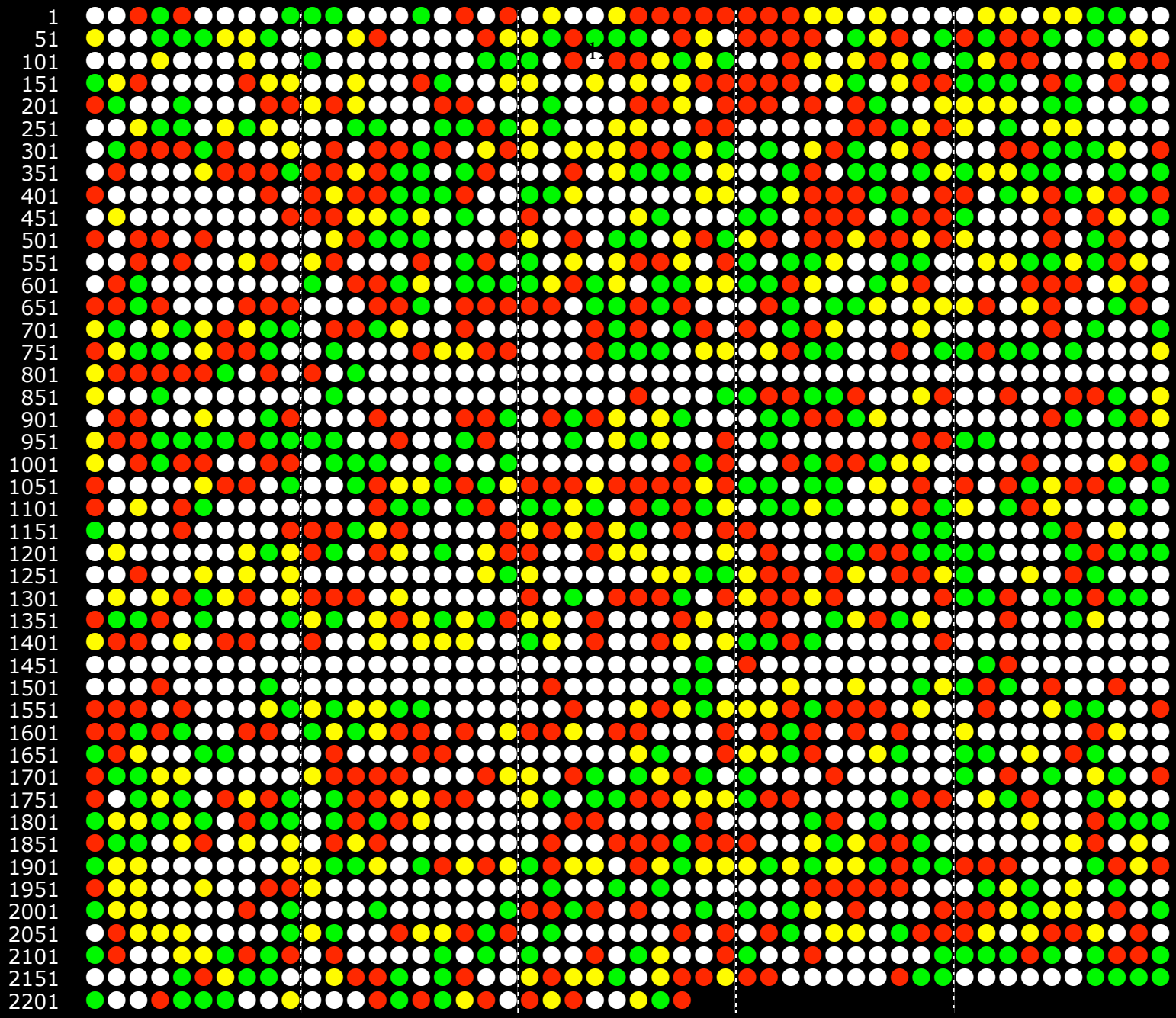
Protein	Log2 ScRatio2	*q*-Val	Log2 ScRatio4	SD	n	Log2 PiRatio2	*q*-Val	Log2 PiRatio4	SD	n	Expression Direction
P13793|FMA_PORGI	-2.25	0	-1.82	0.90	8	-2.25	0	-1.87	1.85	15	↓
PG0121	1.84	0	2.78	1.80	4	1.09	1.71E-04	5.33	4.11	5	↑
PG0293	1.83	0	1.96	1.36	10	1.79	6.80E-14	2.53	1.84	16	↑
PG0377	1.38	0	2.01	1.68	11	1.99	3.10E-15	3.05	2.45	19	↑
PG0378	2.24	0	2.47	1.16	8	2.19	1.93E-10	4.05	2.70	17	↑
PG0386	2.36	7.69E-12	2.16	0.00	2	3.28	6.11E-07	4.39	2.58	5	↑
PG0449	1.53	0	1.67	1.39	7	2.03	1.80E-13	3.44	3.03	26	↑
PG0776	1.23	2.10E-13	1.13	0.72	4	1.59	1.02E-07	2.33	2.29	17	↑
PG1006	-4.32	0	-3.12	1.75	4	-5.34	0	-4.70	1.42	28	↓
PG1082	2.05	0	1.49	0.54	6	1.56	1.93E-07	2.95	2.28	17	↑
PG1085	2.45	0	3.10	1.51	3	1.40	8.18E-03	6.92	4.42	5	↑
PG1279	2.01	0	2.22	1.70	10	2.37	0	4.35	3.00	17	↑
PG1341	1.84	0	2.03	1.66	5	2.09	5.19E-09	3.59	2.52	10	↑
PG1551	-0.74	8.80E-08	-0.63	0.23	5	-0.73	1.23E-03	-0.70	0.54	8	↓
PG1729	1.28	0	1.11	0.21	6	1.35	1.29E-07	1.89	1.43	10	↑
PG1809	1.90	0	1.82	0.64	6	2.03	4.58E-14	2.26	0.66	7	↑
PG1928	2.69	0	2.82	0.65	3	1.43	1.65E-04	3.26	2.16	4	↑
PG1940	2.94	0	2.86	1.23	19	2.14	0	5.50	2.77	38	↑
PG2130	-6.64	0	-2.97	0.50	6	-6.92	2.78E-07	-5.27	2.92	11	↓
PG2168	-4.06	0	-2.36	2.22	3	-6.19	5.46E-09	-3.98	3.38	12	↓

Key to Supplemental Figures S1-S8

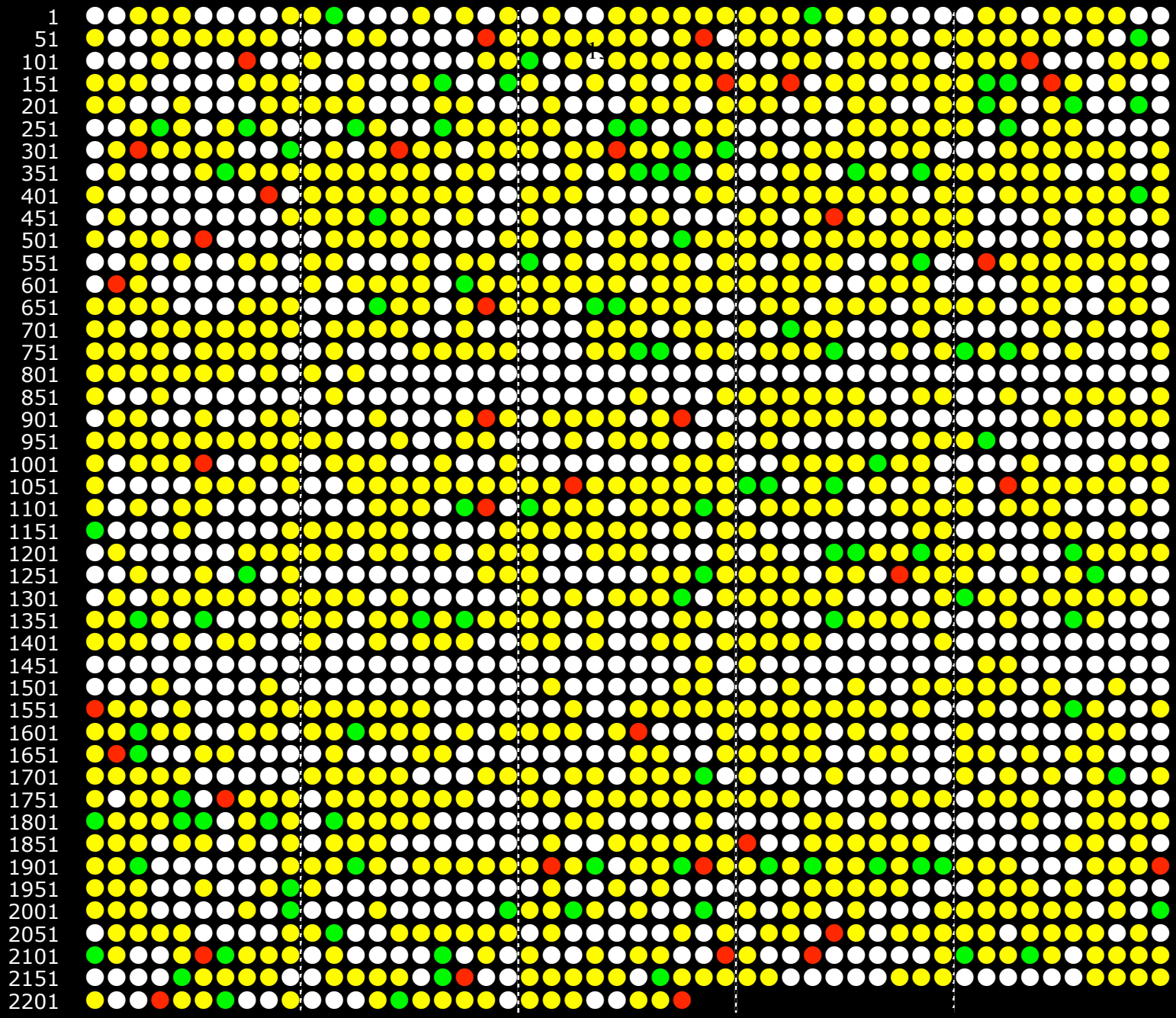
Each false color representation shows the entire *P. gingivalis* proteome in the order assigned based on the W83 annotation by TIGR. References for the annotation are given in the main text, Sections 1 and 2.2. The exception is the ATCC 33277 FimA protein sequence, the last ORF at position 2228 in the lower right portion of each graph. Each plot represents one of the eight experimental conditions shown in Table 1 of the main text and described in Section 1. Each circle represents an ORF: green indicates over-expression of internalized *P. gingivalis* relative to a control population, red indicates under-expression, yellow indicates qualitative detection without evidence for expression change, and white indicates qualitative non-detects.



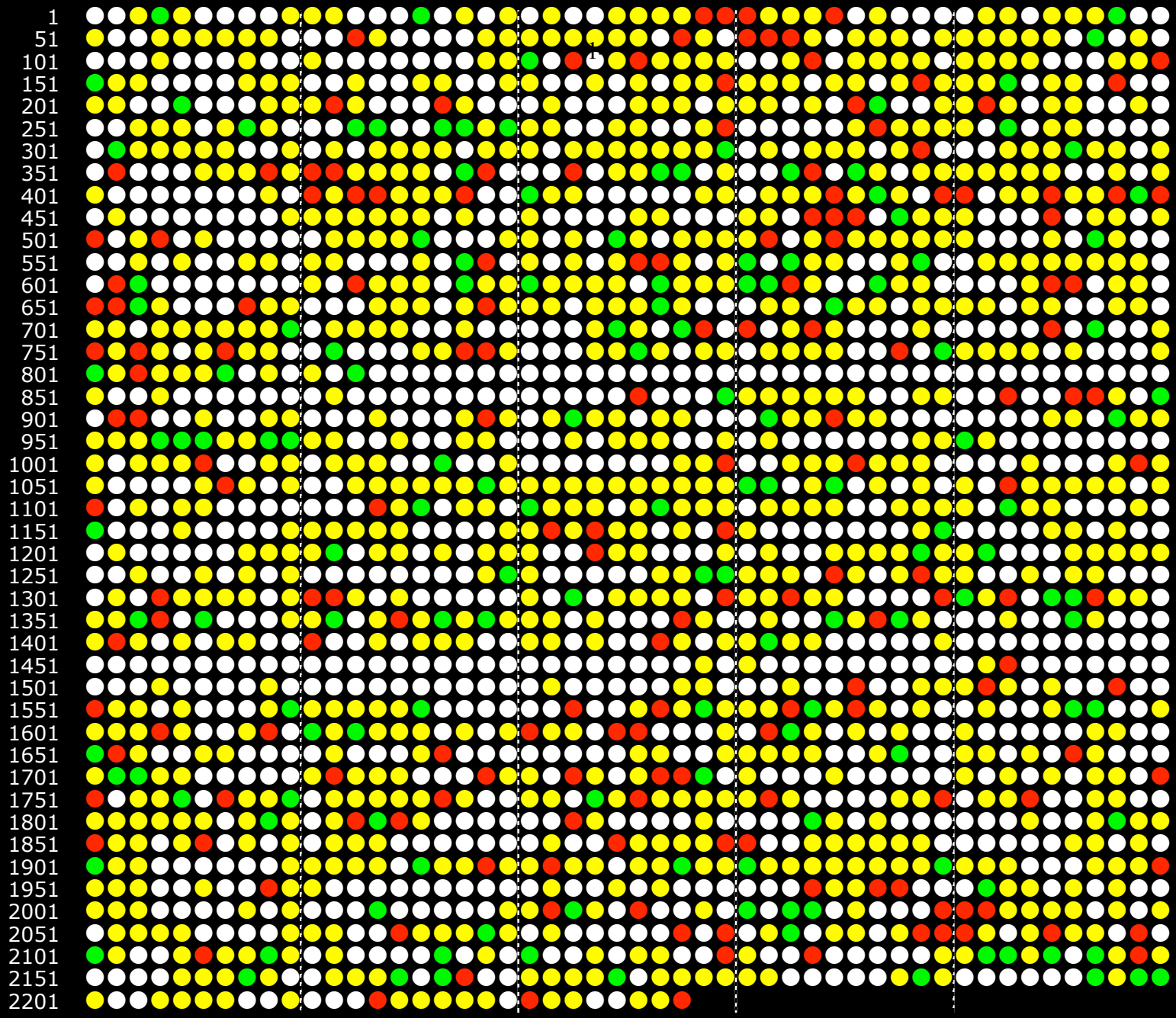
ScRatio1: PG_PP/PG_normal ratio calculated by protein level spectral count method



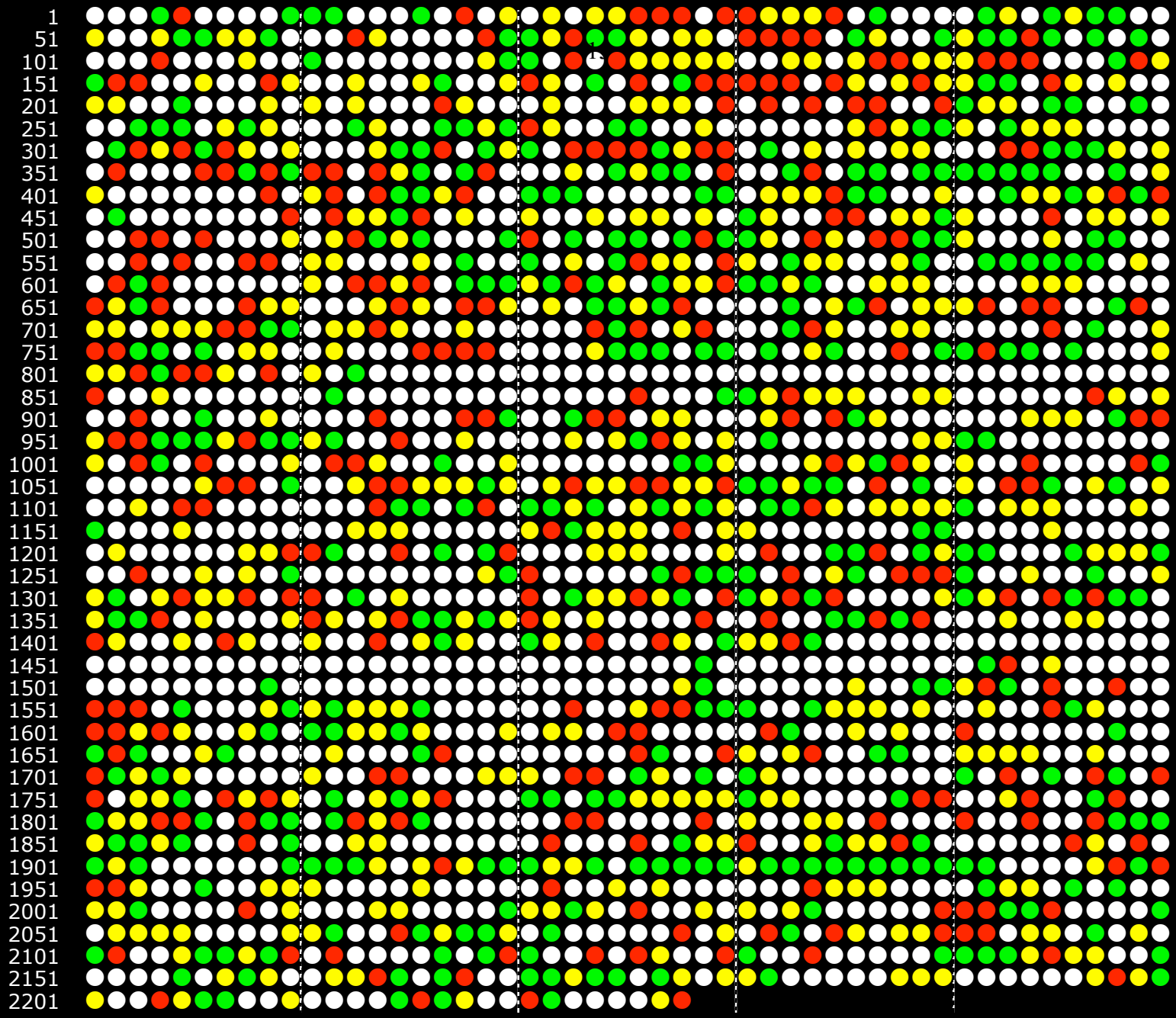
PiRatio1: PG_PP/PG_normal ratio calculated by protein level peptide intensity method



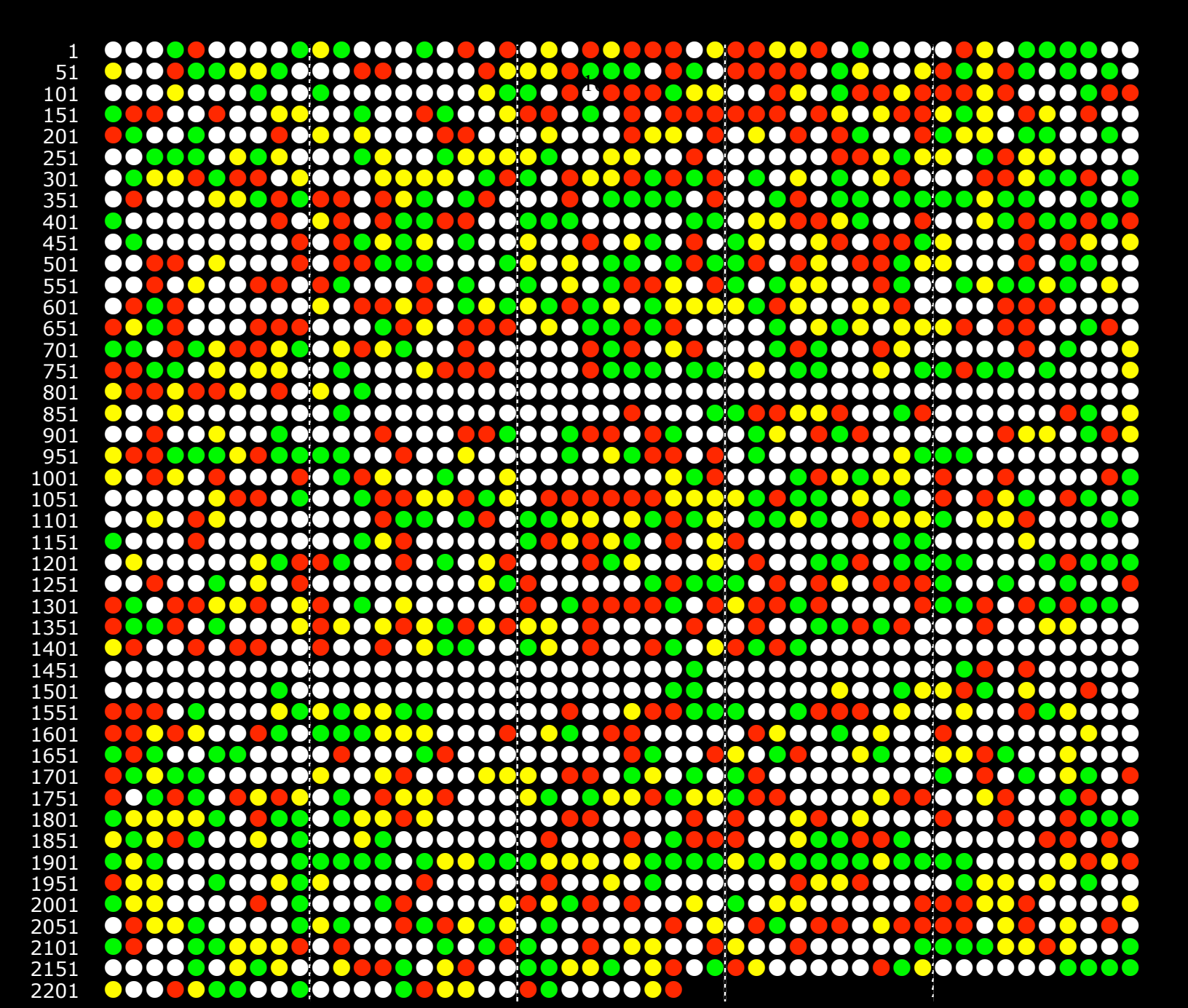
ScRatio3: PG_PP/PG_normal ratio calculated by peptide level spectral count method



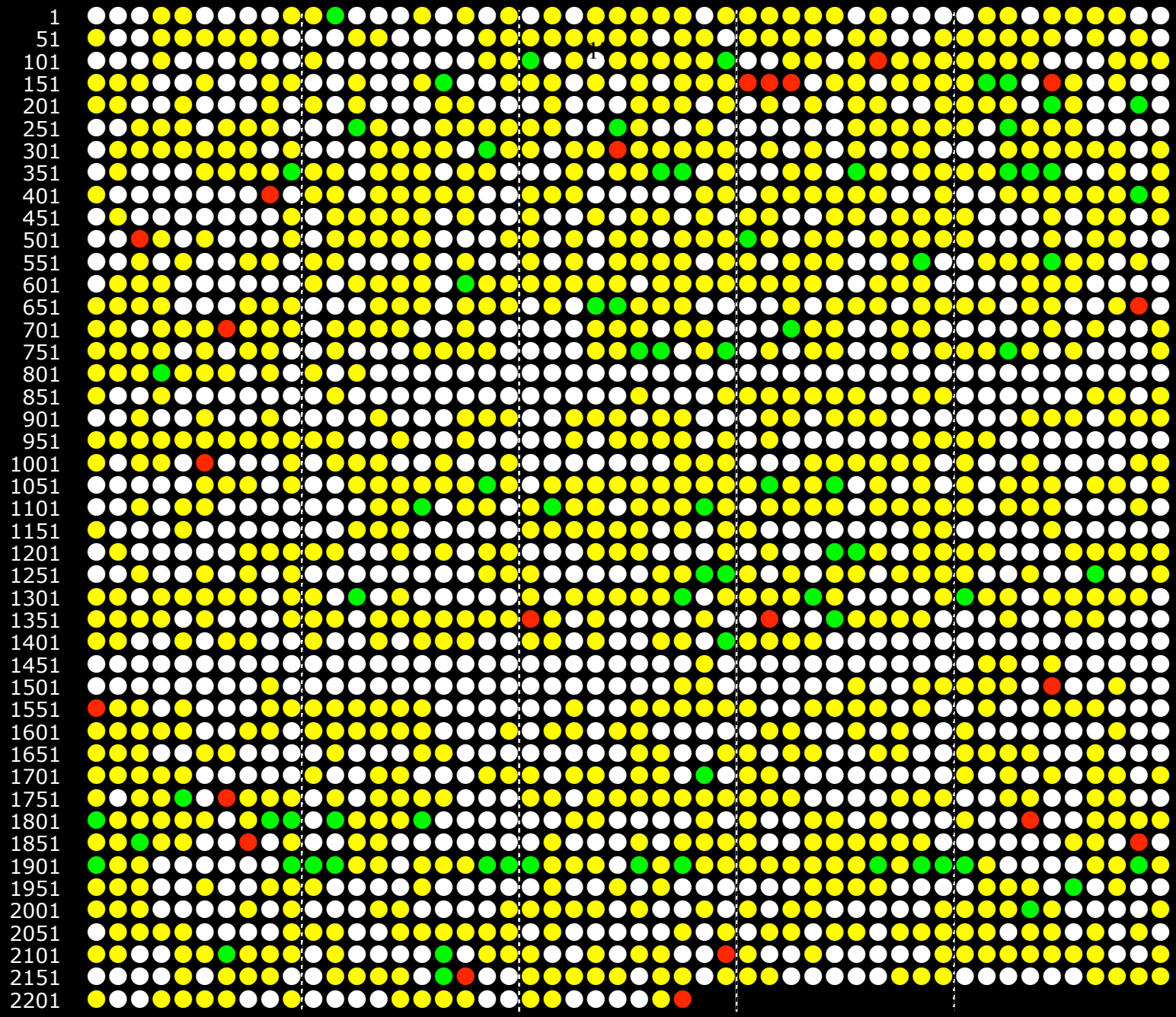
PiRatio3: PG_PP/PG_normal ratio calculated by peptide level peptide intensity method



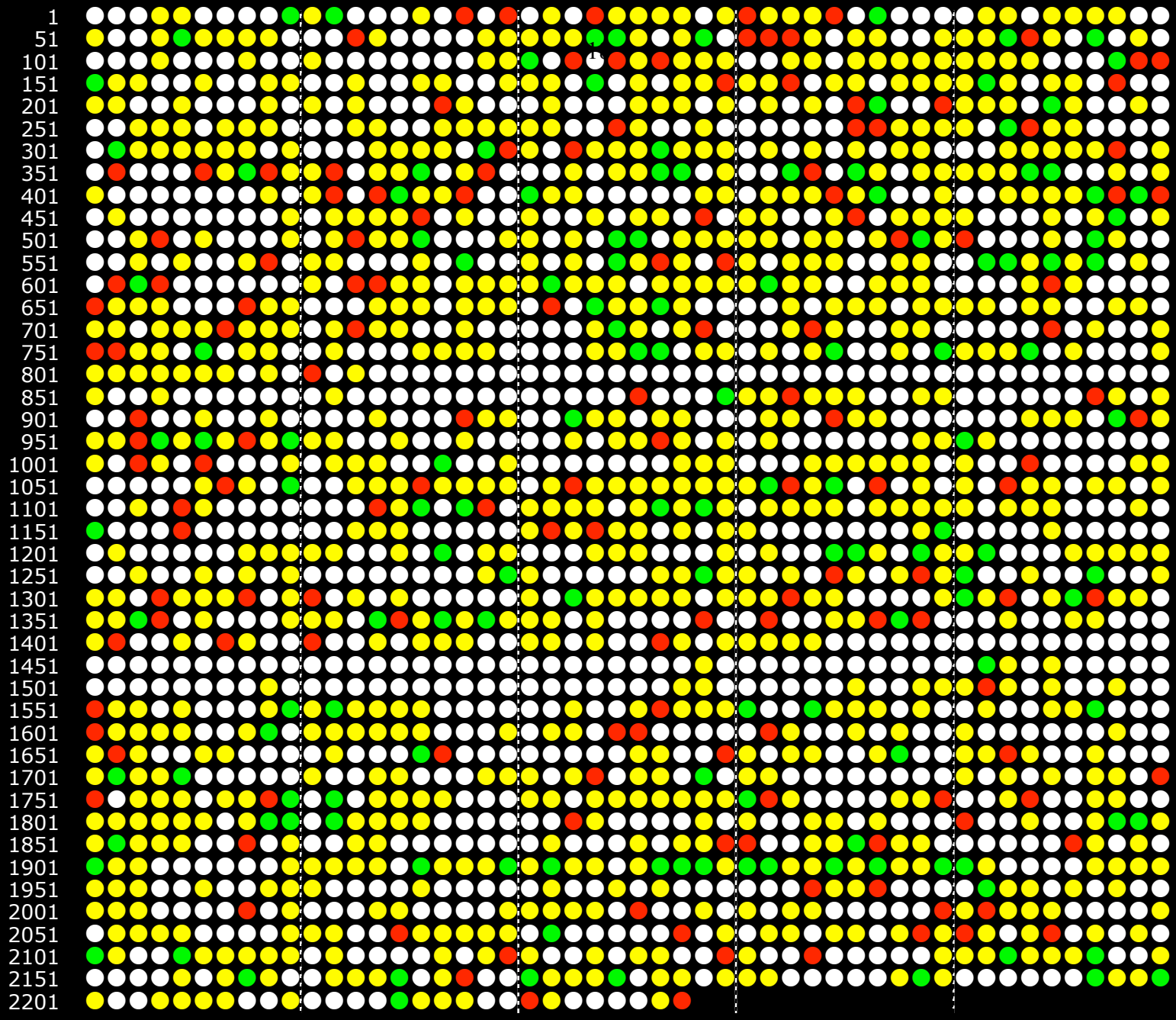
ScRatio2: PG_PP/PG_PPC ratio calculated by protein level spectral count method



PiRatio2: PG_PP/PG_PPC ratio calculated by protein level peptide intensity method



ScRatio 4: PG_PP/PG_PPC ratio calculated by peptide level spectral count method



PiRatio4: PG_PP/PG_PPC ratio calculated by peptide level peptide intensity method.